# A comprehensive review on the application of artificial intelligence in drug discovery

Ashrulochan Sahoo[1] and Ghulam Mehdi Dar[2*]

[1]*Department of Pharmaceutical Sciences and Natural Products, Central University of Punjab, Bathinda-151401, Punjab, India*
[2]*Department of Biochemistry, Govind Ballabh Pant Institute of Post-graduate Medical Education and Research, Jawahar Lal Nehru Marg, Raj Ghat, New Delhi-110002, Delhi, India.*
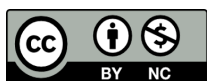
Abstract: The 21st century is witnessing immense achievements in human history, starting from home science to space science. Artificial Intelligence (AI) is a salient one among these feats, the critical factor of the 4th industrial revolution. Health is the primary and essential asset for the continuity of human civilization on this planet. Not only must we address the deadly existing diseases like Cancer, AIDS, Alzheimer's, heart diseases, gastrointestinal diseases, etc., but on top of that, we must effectively predict, prevent and respond to potential pathogens capable of causing havoc like the recent outbreak caused by SARS-CoV-2. AI-enabled technology with the computational capacity of a computer and reasoning ability of humans saves surplus labor and time that is majorly consumed in target validation, lead optimization, molecular representation, and designing reaction pathways, which traditionally is a decade-long way of searching, visualizing, studying, imagining, experimenting and maintaining a ton of data. This article would focus on how AI will help find the drug-like properties in the compound screening phase predicting the Structure-Activity Relationship (SAR) and ADMET properties in lead identification and optimization phases, sustainable development of chemicals in the synthesis phases up to AI's assistance in the successful conduct of clinical trials and repurposing.

Keywords: artificial intelligence; computer-aided drug design; deep learning; drug discovery; drug repurposing; healthcare; machine learning

Abbreviations

AAE	Adversarial autoencoder	**LSTM**	Long-short-term memory
AI	Artificial Intelligence	**MDCK**	Madin-Derby canin kidney
ANN	Artificial Neural Network	**PAMPA**	Parallel artificial membrane permeability assay
CADD	Computer-Aided Drug Discovery	**PCA**	Principle component unit
CASP	Computer-assisted Synthesis Planning	**QSAR**	Quantitative structure active relationship
CNN	Convolutional Neural Network	**R&D**	Research and Development
DBN	Deep Belief Neural Network	**RBM**	Restricted Boltzmann Machine
DD	Drug Discovery	**RNN**	Recurrent Neural networks
DNN	Deep Neural Networks	**SMILES**	Simplified molecular-input line-entry specification
GAN	Generative adversarial network	**SVM**	Support vector machine
GCN	Graph convolution network	**TPU**	Tensor processing unit
GPU	Graphics Processing Unit	**VAE**	Variational autoencoder
GraphMem	Graph Memory network	**XAI**	Explainable Artificial intelligence
LBVS	Ligand-based virtual screening		

Mr. Ghulam Mehdi Dar
Department of Biochemistry,
Govind Ballabh Pant Institute of Post-graduate Medical Education and Research, Jawahar Lal Nehru Marg, Raj Ghat, New Delhi-110002
Delhi, India
E-mail: gmehdi9k@gmail.com

1. Introduction

Successfully introduce a drug to the mass is like finding a needle in the haystack, something very different from precise engineering. The process of developing a new drug is not that straightforward as it seems to the normal consumer. It has its turns and twists; even a small casualty can be fatal to a large group of people. Hence, drug discovery is a very cautious process [1]. With each passing

day, the perimeter of chemical library data, enzyme and proteomic data, genomic sequencing data, the overall biomedical data is increasing sharply. In such a vast ocean of data, finding one accurate molecule goes hand to hand with many known and unknown opportunities. Hence, the process is definitely non-linear with multiple arbitrary effects and results and very fragile to errors. The contribution of such vast amounts of factors makes the discovery of a new drug very tough and expensive in the human workspace; this inefficiency leads us to the

chemical engineering, biological science, and the active utility of computers lead to the foundation of modern drug discovery. From serendipitous drug discovery to Computer-Aided Drug Discovery (CADD), medicinal chemistry has come a long way via peaks and valleys. Though the past years have seen many hardships to bring a new drug to market through many hurdles, it has been reckoned that in this decade, the use of AI will bring out the best of ever [12, 13]. **Figure 1** presents a brief timeline of advancement in the field of AI.
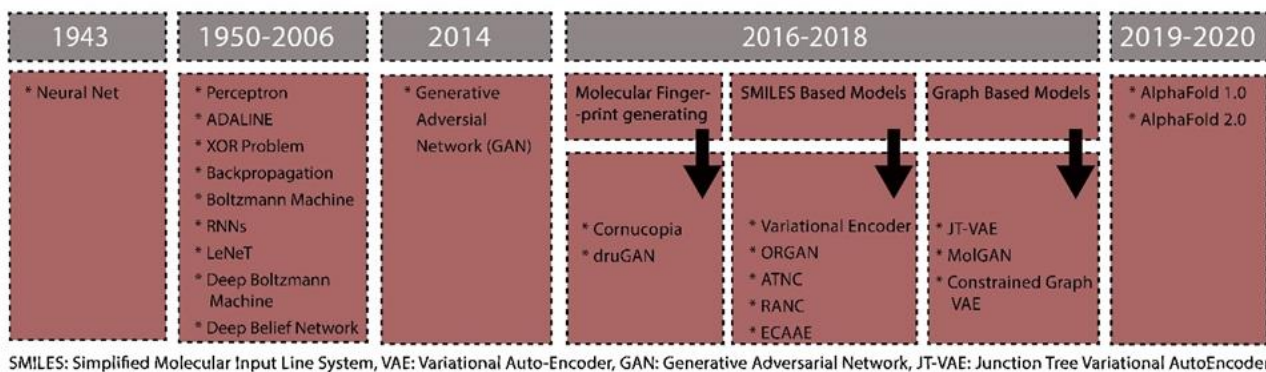

Figure 1. A brief timeline of AI development

accumulation of an external factor called Artificial Intelligence [2].

The twenty-first century has noticed the dawn of a new era of human revolution: the application of AI in daily lives. In 1997 IBM's Deep Blue, a chess-playing computer, was the first to beat world chess champion Garry Kasparov, and in 2017, Google's DeepMind programmed AlphaGo, a self-learning Go player AI, which knocked out the world's strongest Go player Lee Sedol [3, 4]. From smartphone assistance like Google assistant, Apple's Siri, Amazon's Alexa, image and video processing, music and video recommendation to self-driving cars like Tesla and Waymo (by Google) [5, 6]; per diem AI is providing enormous help and support. In fact, AI is the key factor of the fourth Industrial Revolution. AI provides a wide road to explore and hence is currently being in adoption to every discipline of study [7]. AI, along with its two-subfield machine learning (ML) and Deep Learning (DL), is now vastly used in pharmaceutical sectors for pharmacophore modeling, molecular designing, pharmacological data analysis, assay analysis, chemical synthesis, and drug trial monitoring [8-11].

This paper sheds some light on the problems the pharmaceutical sector is facing in fast-track drug discovery and an adventurous journey of AI to alleviate those complications and discuss their forthcoming. The literature is currently more or less focused on applying AI in a particular phase of drug discovery; however, here, we aim to provide a cumulative update covering all the stages of drug designing and discovery.

2. Drug discovery and AI

In the past three decades, the drug discovery process has been evolved in many folds. Immense advancements in

2.1. Approaches in drug discovery

The major shift in molecular drug discovery is dated back to the 1990s, where large numbers of molecular libraries were established through target-based High-throughput screening HTS using combinatorial chemistry [14]. Combinatorial chemistry is the synthesis of thousands of compounds in one single process [15]. The HTS is an automatic process where a bunch of compounds could be rapidly screened based on their receptor-molecule interaction [16]. The advancements in new disciplines of science such as computational drug design, biotechnology, molecular biology, and medicine; human genetics enhanced the knowledge of transgenic animal models, the discovery of new drug targets (proteins, enzymes, and receptors), biomarkers, and understanding of new and existing diseases and their mechanisms. The more the study of molecular library data, proteomics, and genomic data evolved, the sharper the understanding became. Combining all the available data sets, different hypothesizes had been proposed to ease the process of DD. With the ADMET properties: absorption, distribution, metabolism, excretion, and toxicity data, quantitative structure-activity relationship (QSAR) data, receptor-target interaction data, finally, the computer-aided drug discovery (CADD) came to exist by using computer simulations known as *in silico* drug design [17]. **Figure 2** illustrates the different phases and techniques that generate the relevant data for drug discovery.

The process of drug discovery can be broadly classified into four major stages: (i) target recognition, (ii) target development, (iii) pre-clinical studies, and (iv) clinical studies. Every stage contains many layers of experiments to bring the best fit. Each stage is interdependent to the other [18]. We will go through these stages with respect to the ML approach.
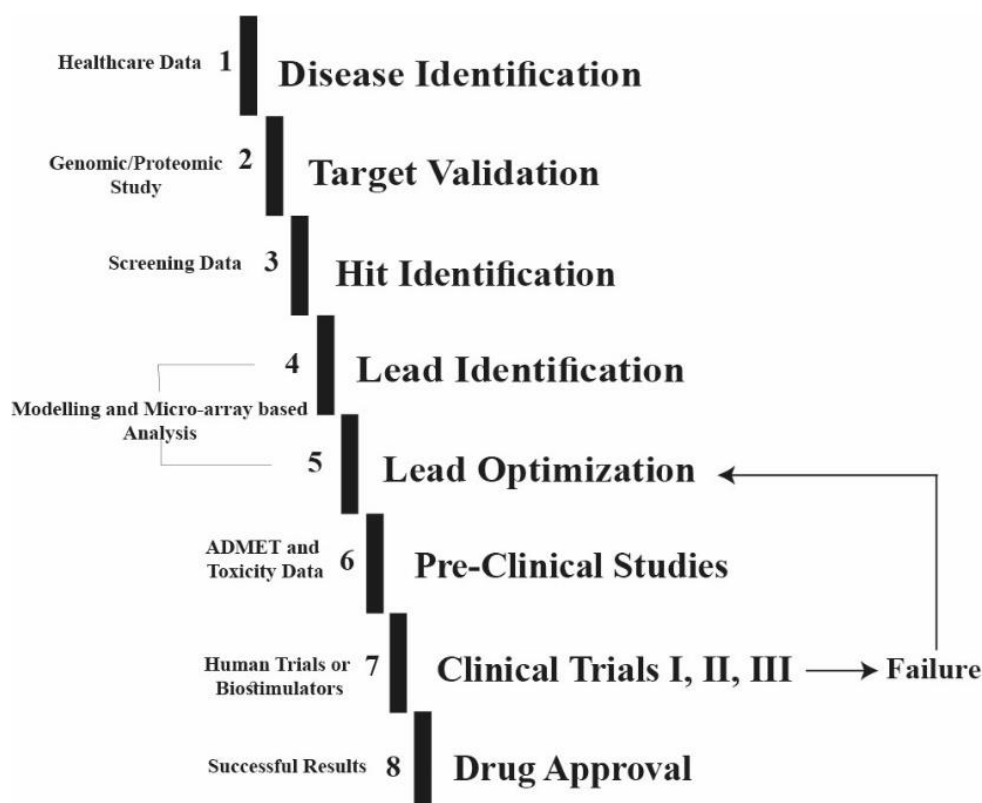
Figure 2. Phases of drug discovery using data from different sources [19].

2.2. Need for the drug discovery

Health is the primary asset to a prosperous civilization [20]. As fast as humans are adding revolutionary chapters to civilization, knowingly and unknowingly, some situations are compromised. This then leads to severe disasters, *i.e.*, the recent SARS-2 COVID-19 pandemic. Not only

lymphatic filariasis, dengue, guinea, etc. [20, 22]. Though we have successfully controlled some deadly diseases of past decades, some new threats and diseases are still coming on the way; and we should be prepared. Drug discovery that includes either novel drug synthesis or drug repurposing remains the two means of our defense against such potential diseases (**figure 3**).
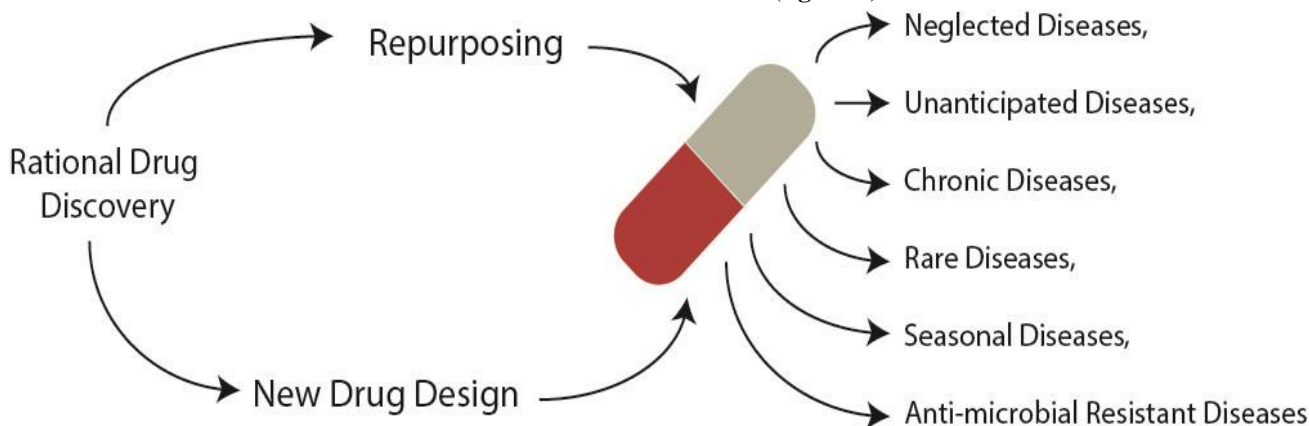

Figure 3. The need for new drugs; they can be new drugs by rational drug development or repurposed drugs.

COVID-19 but also the existing chronic diseases like cancer, diabetes, gastroesophageal reflux disease (GERD), various cardiovascular diseases, epilepsy, acquired immune deficiency syndrome caused by Human immunodeficiency virus (HIV-AIDS) infection, is nowadays taking serious turns in every group of the population worldwide [21]. Rare diseases like thalassemia, hemophilia and primary immune deficiency diseases in children are also serious issues. To also address tropical diseases like leishmaniasis, leprosy,

2.3. Problems with conventional approaches of drug discovery

The huge advancements in technology and managerial approaches in the research and development R&D sector exceptionally uplifted modern drug discovery operations with new diversifying perspectives. Albeit, the drug to market ratio is very drastic, falling around 80-fold (in terms of inflammation-adjustment) in comparison to the 1970s

drug approval rate due to its complexity, longevity, and expensive nature [23]. The process of drug discovery goes through many experiments and researches done by a variety of professionals. Every year, pharmaceutical companies invest money to develop new drugs for diverse diseases; fortunately, one might get a chance to the market out of hundred projects. It takes approximately 12 years with a payout of US$3 billion and lots of manpower to bring out a new drug candidate. The fate of those unsucceeded projects is termed as "R&D inefficiency of the pharmaceutical sector" [24]. This R&D inefficiency of pharmaceutical companies also depends on many factors, *i.e.*, geographical location, the criticality of diseases, market regulation policies, availability of active pharmaceutical Ingredients APIs, etc. According to Scannell *et al.*, the R&D inefficiency is because of the four factors: (i) the 'better than Beatles' problem refers to the resistance by endorsed drug molecules to the upcoming new drugs by setting approval, adoption, and reimbursement barriers; (ii) the 'cautious regulator' problem refers to the constant upgradations to the drug safety regulations by respective authorities; (iii) the 'throw money at it' tendency is the investment of companies in other sectors by downsizing R&D sector to top the market competition; (iv) 'basic-research-brute-force' bias points out the proneness to overestimate the ability of advances in basic drug discovery approaches [25]. Due to these inefficiencies' consumer faces hefty pricing of medicines. They pay both for the succeeding drugs as well as for those failure projects. With this, even some tropical diseases on their primary level were left underrated and unaware, which gradually became notorious to a group of people [24, 26].

Insufficiency in advanced biomolecular tools such as chemical probes and antibodies is also an important setback in molecular drug discovery. In-depth biological understanding is limited to less numbers of proteins. One in three proteins remains understudied; their function in human biology and role in disease studies remains an enigma. To date there's only 11% of human proteome has been explicated; this is recognized as a causality dilemma, which keeps a major portion of proteomic and genomic studies in shadows. Ultimately, this causality dilemma slows down the progress of modern drug discovery [27].

Like a tree, science is growing, and with each passing decade, new disciplines have been emerging from it like branches. Knowledge and experience within a particular field are also increasing swiftly; in fact, thousands of new articles have been added to each branch of science every year. Every year, the MEDLINE data (repositories of medical knowledge) increases around eight hundred thousand plus. The ZINC library data (free database of compounds for VS) has also seen a peak of thousand-folds between 2005-2019 from 700,000 entries to 1.3 billion entries. Accordingly, the pharmacological data, protein data bank PDB entries, in vitro HTS data, molecular drug design data, experimental chemistry data, and toxicology data are also increasing. Each stage of drug R&D involves data mining and analysis to create a hypothesis and experimentally testing them. This data, as a whole, is getting very complicated. In a rapid manner, the human brain individually or in a team is less capable to create and process such amount of multivariable complex hypothesis at million data points flawlessly. It is also exhaustive and time-consuming for humans to monitor every stage of DD. There also exists an experience gap between experts of different subfields; that affects DD in many ways [28]. **Figure 4** illustrates the contribution of different scientific disciplines in supplying the necessary data for the molecular drug designing processes.
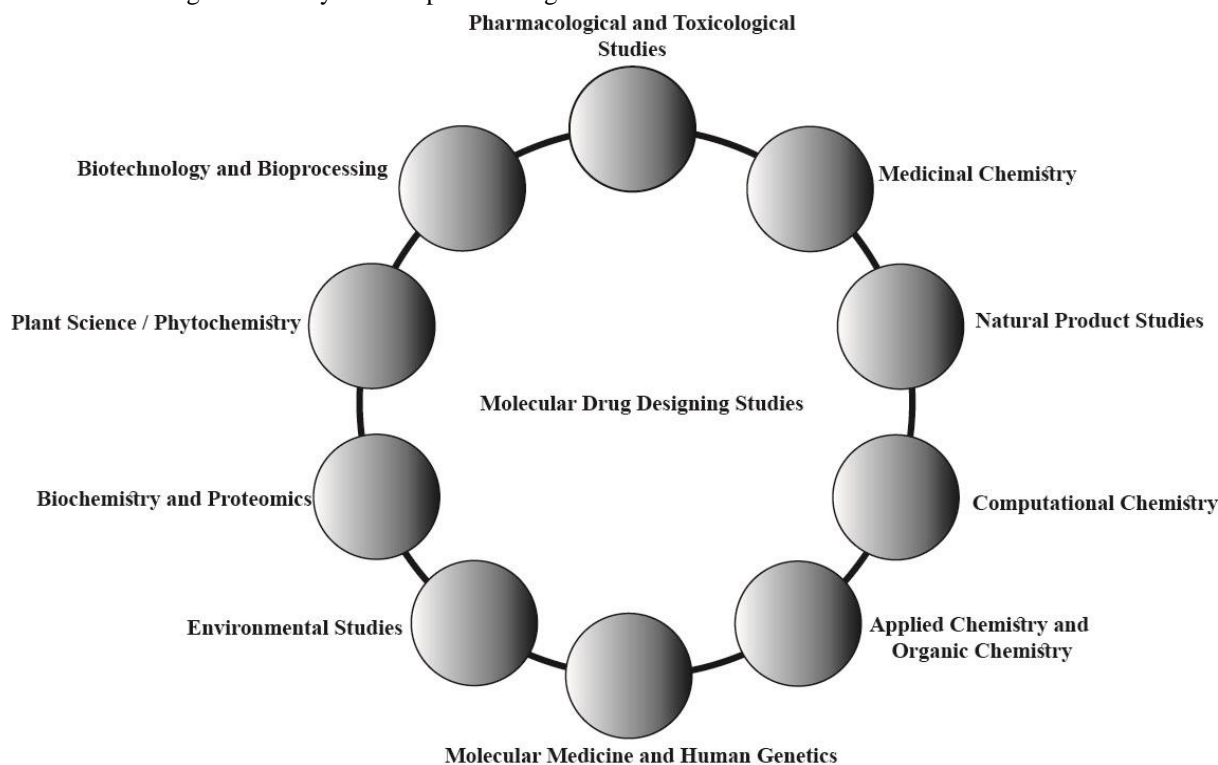

Figure 4. Data mining for drug discovery from different fields of science.

From target identification to clinical trials and approval, these are the current setbacks in pharmaceutical sectors retarding the progress of new medicines. However, the approach of expert-driven study backed up by a data-driven study in R&D methods has reassured breakthroughs. The adoption of AI in the drug discovery process has given a ray of hope. The evolution of computational tools has proved to be an efficient, cost-effective alternative to conventional drug discovery approaches (**table 1**) [29]. Gisbert *et al.* demonstrated the humongous application of ML approaches in chemocentric and molecular informatics studies in three steps: (i) selection of problem-specific descriptor sets to find out the essential properties of involved molecules; (ii) molecular property driven scoring or metric schemes to compare the encoded molecules; (iii) implementation of suitable ML algorithms to identify

studies, *i.e.*, early QSAR and drug-likeness property studies of active molecules, Hansch *et al.* used computers in 1964. Then, various pattern recognition methods have been used to study the specificity of active molecules. Around the 1980s, the implementation of artificial neural networks (ANN) in computational studies helped elevate regression tasks. Perceptron, Neocognitron, were prototypes of ANN used in those early DD studies. For the first time in 1989, Qian and Sejnowski *et al.* published the application of neural networks in protein secondary structure prediction. Around the 1990s, fully automated molecular designing models had been introduced for integrated learning and decision-making purposes. Using a little advanced ANN, *i.e.*, the backpropagation neural network (BPNN) and mainstream potential ML algorithms, *i.e.*, the support vector machine (SVM) and random forest (RF), those

Table 1. Comparison between conventional approach of DD and AI-driven DD.

Parameters	Conventional Approach	AI driven Approach
Time	Consuming/ Long process	Conserving
Labor (physical and mental)	Consuming	Conserving
Data	Works good with less data	Works best with much data
Budget	Very expensive	Quite less expensive than conventional approach

exceptional features for qualitative and quantitative separation of active compounds. The use of AI enhances speed and ease of scalability in modern drug discovery approaches [30].

3. AI in rational drug discovery: a paradigm shift

A notable AI researcher, Antonio *et al.*, has penned down the AI into two parts: weak AI and strong AI. Weak AI is the artificial narrow intelligence (ANI) machine, and strong AI refers to artificial general networks AGI. Weak AIs have been designed to perform some specific operations, *i.e.*, smartphone assistants, smart home assistants, etc., have gained mass acceptance in everyday life. Whereas the strong AIs are meant to play crucial roles in bigger aspects of human civilization, *i.e.*, fully automatic robots to perform surgery, fight in the war, and other complex studies. The popularity of AI's is because of their work efficiency and self-learning ability in any environment; ease of life is just one click away. Even some software companies have started making AI operating system applications to accompany lonely and stressed people; who listen to them, talk to them, and perform tasks for them by gradually learning about their choices and lifestyle. In drug discovery, both weak and strong AIs will assist chemical and biological scientists in each stage of molecular drug discovery. The field of strong AI or AGI is under the show yet; albeit, pharmaceutical companies are largely acquiring ANIs to truncate expense, waste, labor, and failure rate [31].

The application of AI in molecular drug discovery studies dated back to the 1960s. For the physicochemical property

models were made capable of self-learning from experience, problem-solving, and habitant to new situations. In 2014, the deep neural network (DNN) studies boomed with creating the Generative Adversarial Network (GAN) architecture. GAN helped scientists to build molecular architecture with unprecedented generative capabilities [32]. A 2016 model, Cornucopia - molecular fingerprint interpreter, was widely used by chemists to study structure and reaction mechanisms. Variational auto-encoder (VAE) for converting molecular structures to computer understood language translator simplified molecular-input line system SMILES strings has become crucial for molecular design and synthesis studies [27]. The cascaded approach before the modern deep learning approach by combining a number of different types of ML algorithms [33]. The advancement of AI has reduced a small amount of gap between conventional DD and rational DD approaches [34]. In 2017, the pharmaceutical industry started the partnership with software companies or AI manufacturers. A 2018 McKinsey report predicted a 39% revenue increment in pharmaceutical sectors by AI approach over traditional approaches [35].

4. Progress of AI in rational drug discovery

4.1. Use of AI in ligand-receptor binding affinity prediction

The primary step in drug discovery is finding out the receptor related to a disease and understanding its role in that particular disease mechanism. Based on the receptor's molecular nature, several chemical compounds with drug-

likeness are proposed, and the one with the highest receptor-ligand binding score is assumed to be the hit. Search algorithms and scoring functions SFs are two important features for docking study. The SFs are mathematical functions used to predict receptor-ligand binding affinity, measured as the Pearson correlation coefficient (Rp).

There have been different ML and DL-based AIs are proposed to predict the binding score. Two SVM-based models are SVR-Score and ID-Score. Ballester and Mitchell *et al.* had developed a number of RF-based AIs, *i.e.*, RF-IChem, SFC_{score}^{RF}, X-Score, and B2B Score, to study the receptor-ligand binding affinity. Among these, RF-Score had performed better, encoded into large-scale protein-ligand docking website DockThor (https://www.dockthor.lncc.br/v2/). Again, Ballester *et al.* confirmed the better performance of RF-Score-v3 in comparison to X-Score with respect to 16 classical scoring function sets. RF uses decision trees DTs as base learners. This helps to incorporate the algorithm with much variance and flexibility; this high variance reduces the correlation between trees. Hence, it improves the accuracy of the score prediction in the whole ensemble model. For rescoring purposes, NNScore and CScore ANN-based machines have been developed.

In structure-based drug designing, the tools used for molecular docking operate on diverse sampling algorithms, docking, and simulation methods. These tools also use various scoring parameters and functions to predict the most accurate binding score. The methods and functions depend on the three-dimensional structural features of the ligand, which is evaluated by implementing rotational and translational vectors [36]. The only problem with the above-mentioned ANN, RF, or SVM-based ML machines is their need to represent molecules with fixed-length vectors. The

minimized the limitation, as it is capable of extracting the features directly from the 2-D and 3-D molecular structures. Cang and Wei *et al.* developed a multichannel topological neural network TopologyNet using a topological strategy and a CNN model developed by Ragoza *et al.* While the CNN is used to make a 3D grid for each protein-ligand complex, the topological strategy is used to represent the 3D biomolecular geometry of 1D topological invariants into a reduced-dimensionality formulation. This arrangement occurs without altering the important biological properties of the molecule, and across every grid point, the atom densities are stored.

In 2020, Artem *et al.* had introduced Deep Ducking, a novel DL platform for structure-based drug discovery. The Deep Ducking model combined with the FRED docking program was used to dock 1.36 billion molecules from the ZINC15 library against 12 target proteins; the results were pretty striking. It yielded approximately 6000-fold enrichment of high scoring molecules and a reduction of 100-fold data without any loss of favorably docked entities. The Deep Ducking approach utilizes Keras Python library to build and train the feed-forward DNN model on the basis of QSAR properties and docking scores of subsets of a chemical library [37]. The Deep QSAR model works in an iterative manner to approximate the docking outcome for both processed and unprocessed entries for a large number of molecules and removes the unfavorable molecules in a rapid and accurate manner [38].

Albeit the advancements in different ML and DL-based Ais in receptor-ligand affinity prediction have outperformed the classical method; still, it's a challenging topic in rational drug discovery. Due to some partiality or bias in the scores in every respect, it is hard to depend on one step action and one model (**table 2** compares parametric differences

Table 2. Comparison between different scoring functions (SFs) [34, 39]

Parameters	Physical-based SFs	Empirical SFs	Knowledge-based SFs	Descriptor-based SFs
Examples	GoldScore and SFs in AutoDock 3,4	ChemScore and GlideScore	DrugScore and PMF	ML-based SFs i.e., AGL-Score, TNet-BP, BsN-Score, RF-Score, etc.
Working Principle	Based on physical atomic interactions i.e., Lennard-Jones potential (Van der Waals force) and electrostatic force.	Based on QSAR data	Boltzmann static principle	Machine learning approach
Regression Model	Linear regression	Multiple linear regression	Most of models are additive (Linear) regression	Depending on the algorithm type and neural network used (CNN, RF, SVM, etc.)
Efficiency	Works good with a few numbers of terms	It is suitable for a smaller number of descriptors (i.e., <10 weighted terms)	-	Works better at large number of terms (>100 of descriptors can be exploited)
Interpretability	-	Explanation for individual descriptors could be obtained.	-	It provides no explanation (or 'black box' operation)

development of convolutional neural networks CNNs in DL

between various known classical scoring functions).

The modern docking operations with all the available computational tools can hardly exceed 0.1 billion molecules, leaving a large chemical space inaccessible. To address this incongruity, major steps are: (i) to lower the probability of false ligand pose prediction by using innovative conformational sampling; (ii) to separate the chemical compounds as per their drug-likeness (their ability to be hit/lead/fragment for different diseases) by using precomputed physicochemical parameters and drug-like criteria; (iii) to make those innovative methods user friendly and the separated compound library to be publicly accessed [37, 40].

Prediction of the binding score is the crucial step, and its consequences lead to other studies; hence, it should be carried out with utmost attention. The focus is to build a model which can predict accurate binding score with respect to molecular features and stability, despite the inactiveness of the receptor. Among the above-discussed approaches in scoring prediction, DL algorithms have the potential to work in every range of work environment, which will be a great advantage in the rational drug discovery process. ML technique with Gaussian process along with quantum effects and biophysics will also be useful in this regard. Success in protein-ligand docking will raise the curtain from the understudied protein-protein interaction phenomena [41].

4.2. Use of AI in de novo small molecular drug design

Drug design refers to the molecular arrangements and rearrangements to the obtained hit/lead/fragments. This design is done in a precise manner with respect to the availability of chemicals, with the accurate set of desired interacting groups, for the proposed biological functions, and with an eye to the intellectual property rights IPR and standard safety parameters. The early drug design approaches were structure-based. Most of those drugs were prone to synthetic infeasibility, poor drug metabolism, and less minimized toxicity. The recent de novo drug design

Before the use of ML/DL methods, the de novo drug design was completely knowledge-based. Using deep generative networks like convolutional neural network (CNN), recurrent neural networks (RNN), adversarial autoencoder (AAE), and variational autoencoder (VAE) are the most used full models in de novo drug design. These models can even perform better with no preliminary chemical knowledge [45, 46].

The universal scheme to use DL in de novo drug design is:

(i)   To build a model with deep generative networks of one type or a mixed type.

(ii)   To train the model with various sets of reference chemical compounds (from ChEMBL, DrugBank, ChEBI, GDB-17, or FDB-17) via SMILES strings.

(iii) To make the model experienced through coherent training at different biological, chemical, pharmacological data points.

(iv)   To let the model apply the experienced knowledge to arrange the molecule as per the commanded desired properties.

4.2.1. RNN-based

The early studies on deep generative models gained an utmost advantage from the use of RNN to template sets and novel scaffolds [47, 48]. Segler *et al.* demonstrated the autodidactism of RNNs from the trained data to represent molecules as SMILES (**figure 5**); that learned the grammars to valid SMILES representation and generated chemical molecules of different scaffolds and similar properties [49]. Yuan *et al.* reported a new library generation method using character-level recurrent neural networks char-RNN, known as Machine-based Identification of Molecules Inside Characterized Space MIMICS. In MIMICS, the char-RNN was trained to learn the notable features in SMILES strings for the given set of chemicals; thus, it can eliminate
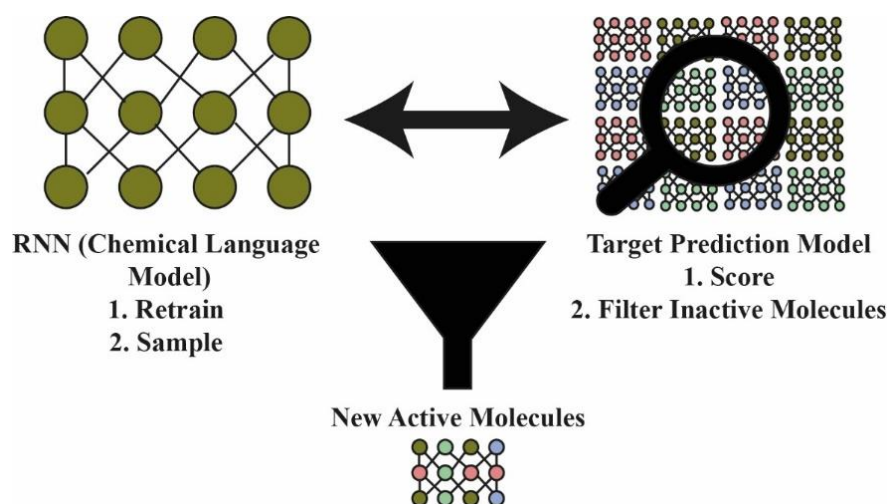

Figure 5. RNN based models in drug design. The RNN model learns the features of desired chemicals from SMILES strings and filters the active molecules from inactive compounds [50].

approaches are ligand-based. Another approach is called the 'inverse QSAR' approach [41-44].

molecules with unwanted properties. In 2018, Popova *et al.* used the stack-augmented recurrent neural network stack-

RNN (extension of RNN architecture implemented with persistent memory unit) to generate a library of novel ligands against Janus protein kinase 2 JAK2 (a non-receptor tyrosine kinase) [50].

4.2.2. LSTM-based

Long-short-term memory (LSTM) networks were being in use for the novel drug design purpose. Hochreiter and Schmidhuber introduced the LSTM architecture to address the vanishing and exploding gradient problems of DNN [51]. The hidden units in the LSTM network are able to choose to be either linear or non-linear through a multiplicative gating mechanism. The linear units help the network potentially count and store a finite amount of information for a long period of time. Hence, LSTM is capable of learning context-free and context-sensitive grammars [52, 53]. In a 2019 article, Awal *et al.* demonstrated the use of LSTM generative neural networks to generate new drug analogs of the reference molecules of the FDB-17 database. For the training of that model, a transfer learning algorithm was used. For hit-lead optimization, Gupta *et al.* combined the LSTM network with RNN and trained the model on ChEMBL with a transfer learning algorithm to learn SMILES grammar. The model performed better optimization of hit-to-lead even with less data [47].

4.2.3. AAE and VAE based

Kadurin *et al.* used a seven-layer AAE architecture to generate molecular fingerprints in cancer drug discovery with definite restrictions. This AAE model was able to identify new molecular fingerprints as per the preset anti-cancer cell properties. This experiment was introductory to AAE architecture in de novo drug design. With the success of the experiment, the researcher group developed a drug generative adversarial network (druGAN) using AAE architecture. In the field of anti-cancer drug discovery, druGAN will be helpful to propose new molecules in a narrow search space [54].

4.2.4. GAN-based

The advanced deep generative adversarial networks (GANs) have the ability to generate drug-like molecules with desired properties using input SMILES strings [55]. Sanchez-Lengeling *et al.* combined GANs with reinforcement learning and proposed an objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC) architecture. ORGANIC is the first DNN to use GAN. The model has a Discriminator component D powered by discriminative networks to collect the initial distribution data of molecules; a reinforcement component R to provide a quality metric to quantify the desirability of given molecules (if the quality matric is $R(x) \in [0, 1]$ for a given molecule x, 1 is meant to represent the desired shift in properties and 0 an undesired change); and a Generator component G to generate the molecules on the basis of the maximum objective function (a linear combination of the D and R parametrized by a tunable parameter λ). While the GAN generates non-

repetitive, sensible molecules similar to the initial distribution of data, the RL biases this generation towards the maximization of the reward. Though ORGANIC is used to design organic photovoltaics, OLEDs, and flow batteries, it has an adversarial setting problem, *i.e.,* mode-collapse during training [56].

To address this common adversarial setting problem, a research group led by Evgeny Putin, Aspuru-Gudzik, and Zhavoronkov used the GAN paradigm and RL to build an original DNN architecture for de novo small-molecule design. They named the architecture Reinforced Adversarial Neural Computer (RANC). RANC uses a differentiable neural computer (DNC) as a generator; that increases generation capabilities by the addition of an explicit memory bank. RANC trained on the SMILES string generates molecular structures as per (i) matching the distributions of the key chemical descriptors, *e.g.*, MW, logP, etc., and (ii) lengths of the SMILES strings used in training. RANC outperformed ORGANIC with reference to several drug discovery matrics, *i.e.*, the number of unique structures, Muegge criteria, medicinal chemistry filters (MCFs), and high QED scores. RANC promises to save time and labor on de novo drug design [57].

4.2.5. Limitation of SMILES and rise of molecular graph approach

SMILES is a great tool to translate molecular structures for computer appreciation and investigation. Despite their wide application in computational chemistry, their use is always not that much revealing. The main complications with the use of SMILES often observed are (i) generation of non-reasonable molecules; and (ii) conjuring of single-character-based molecules [34]. The use of a molecular graph, *i.e.,* graph memory network GraphMem in the development of generative molecules, is an alternative approach for de novo molecule design. To represent and generate molecules by portraying them with a molecular graph is a robust technique; even if some molecular graphs are partially generated, they can be considered subsets of molecules and can be used in chemical checks. Molecular graph generators are trained by VAE or GAN architecture. Currently developed GAN-based molecular graph-generators are MolGAN [58] and Mol-CycleGAN used to generate optimized molecules with much likeness to original compounds [59]. DeepGraphMolGen, a 2020 model, uses graph convolution network GCN and reinforcement learning approach to generate molecules based on drug-like properties and synthetic availability [60].

4.3. Use of AI to predict pharmacological and physicochemical feature of molecules

In the early drug discovery process, without designing all the molecules and actually observing their interaction *in vivo* or *in vitro* assay, observing them *in silico* models has saved time and expenditure in many folds. Big Data, ML, DL, and quantum chemistry approaches are now successively used in the prediction of physicochemical properties, *i.e.,* lipophilicity (log P, log D), aqueous

solubility (log S), intrinsic permeability prediction, ionization constant, melting point, boiling point, Pharmacological properties like absorption, distribution, metabolism, excretion and toxicity ADMET (**figure 6**) [61-63].

In the AI-driven chemical synthesis process, computer algorithms are trained on (i) starting material selection, (ii) reaction prediction, (iii) reaction condition prediction, (iv) synthetic route planning, and (V) retrosynthetic pathways [71, 72]. The data for machine training is evaluated from
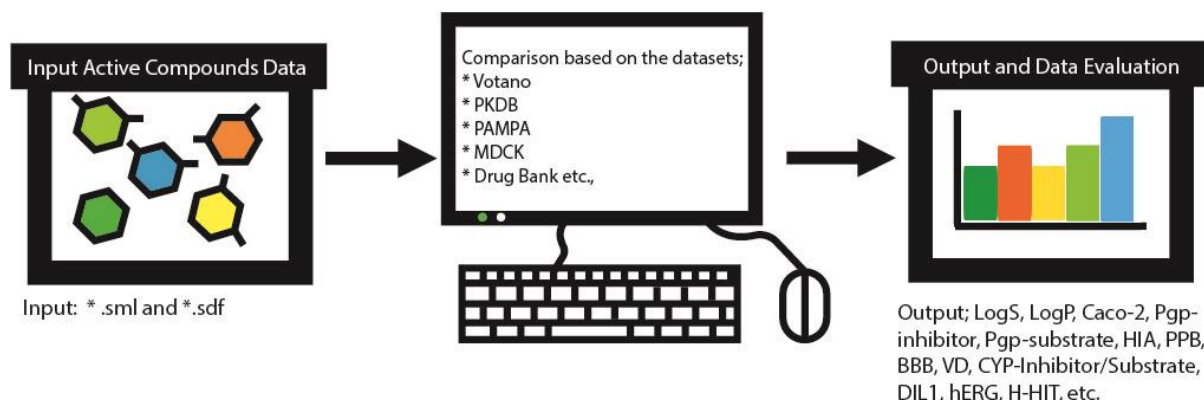

Figure 6. Physicochemical and pharmacological feature prediction by ML approach. The structures designed molecular library is converted to .sdf or .sml format and then imported to the machine. The machine is trained on different data points from various sources, *i.e.*, DrugBank, Votano, PAMPA, etc. The AI processes the data using encoded ANN and exports the ADMET and physiological properties as graphs and charts for comparison purposes [61].

To predict octanol-water partition log P, ALOGPS is a model that uses associative neural networks, the combination of feed-forward network and k-nearest neighbor (*k*NN). Undirected graph recursive neural networks (UG-RNNs) and graph-based CNN are used to predict aqueous solubility [64]. RS-predictor (using hierarchical descriptor and quantum chemical, atom-based descriptor), SMARTCyp and Xenosite (combining ANN with topological, quantum chemical, and SMARTCyp descriptor), CypRules, MetaSite, Metapred, WhichCyp are available tools to predict sites of metabolism [65]. Many ML methods are used for toxicity study, *i.e.*, SVM, relevance vector machine (RVM), regularized-RF, RVM boosting (RVMBoost), SVM boosting (SVMBoost), AdaBoost, and C5.0 trees. DL-AOT, pkCSM (uses graph-based structural signatures), admetSAR, LimTox, and Toxtree web tools and packages are available for toxicity studies in de novo drug design [34].

4.4. Use of AI in de novo chemical synthesis

The synthesis of a new drug is always knowledge-driven. Computer-assisted synthesis planning CASP has transformed rational drug discovery without needing the medicinal chemists to synthesize all the selected molecules by searching tones of organic reaction pathways [66]. A new drug should follow Lipinski's rule of five: (i) molecular weight should be less than 500 Da; (ii) there must be less than five H-bond donor atoms and (iii) H-bond acceptors less than ten; (iv) octanol-water partition coefficient logP should be less than five [67, 68]. Other aspects that should be considered before the beginning of the synthesis are the yield of the reaction with atom economy AE, process mass intensity PMI and costs of materials to be used [69]. Moreover, the reaction reagents, catalysts, products, and byproducts should follow green synthesis norms and other safety parameters [70].

various databases, *i.e.*, Chemical Abstract Service CAS with 127 million reactions (largest provider), Reaxys, SPERSI, Pistachio, and United States Patent Office USPTO. SOPHIA, LHASA, CAMEO, SYNCHEM, EROS, and RASA algorithms are used to design various ML tools [73-75].

4.5. Use of AI in pre-clinical and clinical trials

The synthesized drug has to pass pre-clinical studies in animals to enter the full passage of clinical trials. In phase I of the clinical trial, investigators use a small quantity of drugs on twenty to eighty healthy human volunteers (with no medical conditions) for several months to study human pharmacology and evaluate ideal dosage. Phase II comprises hundreds of infected volunteers (humans with the disease that the new drug is meant to treat) with the same dose for several years to study the interaction and other therapeutic conditions. In phase III, thousands of randomly chosen infected volunteers (up to 3000) are observed for several years. Phase III is a double-blinded trial (both the observer and volunteers don't know what medication they are using) to confirm the findings of the early phase. Here, in phase III, the new drug gets approved; however, its safety and other therapeutic uses are still observed in phase IV [76].

In clinical trial failure rate of proposed drugs is very high due to (i) inefficient volunteer selection; and (ii) inability to effectively monitor the observation [27]. ML and DL approaches have been proposed to prepare the study, regulate required parameters, and constantly monitor trial success rates to address these casualties in a clinical trial. Various AI tools are used to predict human-relevant biomarkers of diseases to recruit a specific patient population in Phase II/III trials [77, 78]. The machine is

designed in such a manner that it notes down every change in the patient's medical condition electronically. IBM Watson uses a DL-based clinical trial matching system to maintain and analyze structured and unstructured electronic medical records of patients to create and select suitable patient profiles [79]. PrOCTOR predicts toxicity probability. AiCure is a mobile application used to monitor phase II clinical trial data of schizophrenia patients; it showed 25% improvement in monitoring data compared to traditional 'modified directly observed therapy' [24].

4.6. Use of AI in drug repurposing

Repurposing approved drugs and under development drugs (failed projects) is a new smart and logical approach in the rational drug discovery process; to defend obscure therapeutic prerequisites of unexpected, rare, and ignored diseases. Repurposing of drugs works because (i) different diseases share molecular pathways and genetic factors, and (ii) drugs have multiple targets. Repurposing needs a lot of data from various perspectives [80, 81]. To feed this,

Table 3. Partnership between pharmaceutical industry and AI industry [24, 35, 93-95].

Pharmaceutical Organization	AI Organization	Purpose/ Focus Area	Approach
Pfizer	IBM Watson Health [94]	To improve immune-oncology studies by identifying new drug targets, combination therapies, and facilitating patient selection strategies	ML, natural language processing and cognitive reasoning
	XtalPi [95]	To predict chemical, mechanical and protein-ligand interaction features of molecules	ML algorithms, cloud computing and quantum mechanics
Roche	OWKIN	Drug discovery, development and clinical trials	ML networks
	Bina	ML-based genome sequencing and personalized medicine	ML based
Novartis	IBM Watson Health	Monitors and Improves health in breast cancer patients	-
	Microsoft	To facilitate research on generative chemistry, gene-based therapies, image segmentation and analysis for smart and personalized delivery of therapies, and to scale the optimization of cell and gene therapies.	-
Merck	Numerate	Novel small molecular lead generation from unnamed cardiovascular disease target.	-
Takeda		Novel drug design and clinical trial for gastroenterology, oncology, and CNS disorder	-
SERVIER		Small molecular modulator design of ryanodine receptor 2 RyR2 (a target in cardiovascular diseases)	Numerate algorithm-driven drug discovery platform
Bayer	Turbine	To predict the response of cancer cells to treatments	-
	Xbird	Tracks real-time data via smartphones, smart watches, etc.	-
	Sensyne Health	Clinical development of new treatments in cardiovascular diseases	Sensyne Health's clinical AI technology platform
	Exscientia	Novel lead structure optimization against cardiovascular and oncological diseases	Exscientia's AI drug discovery platform Centaur Chemist^{TM}

computational techniques are the best suit [82]. The important algorithms used for drug repurposing studies are supervised learning, unsupervised learning, and semi-supervised learning. Supervised model, *i.e.*, DTINet; unsupervised models, *i.e.*, MANTRA, semi-supervised model, *i.e.*, LapRLS and advanced NetLapRLS, LPMIHN, BLM with neighbor-based interaction-profile inferring (BLM-NII), and network consistency-based prediction (NetCBP) method, network-based deepDR are in use for drug repurposing [83-87]. Though these models have promised better performance, still there predictions are not conceived yet [82].

5. Future of AI in drug discovery

AI paradigm has advanced the drug discovery process with new insights as a whole. It has eased the knowledge-based research criteria for chemists, biologists, and scientists. It is boosting the pharmaceutical sectors and their experiments to a new level [88, 89]. The time of the hit and trial approach in drug discovery has shifted to rational drug discovery and development. However, AI has two potential drawbacks (i) data insufficiency and (ii) Black box prediction model. In the current scenario, AI in drug discovery is like the 'blind watch maker'. Though it knows its purpose and mostly accurate, still it's unexplainable; a 'black box' prediction. In science, rational and correct explanation is the main theme. Until the AI becomes explainable XAI, the researchers will guesstimate in shadows. In the near future, with much advancement, the XAI may explain the how. Rapid growth in database libraries has made the data messy as a whole. Also, the search engines are now becoming specific. These two facts are ultimately leading to the possibility of missing in randomness. In contradiction, the DNN machines need a large number of data to be trained upon. The deeper the network, the more data it demands for training and reinforce-learning. This problem can also be addressed by assistantship of AI, which can mine the data and feed the analysis itself [90, 91].

The partnership between pharmaceutical sectors with AI organizations is facilitating research (**table 3**). A number of startups are opening. However, some challenges are still present in the current scenario of rational drug discovery. Peter *et al.* have demonstrated five major challenges; (i) Data governance, (ii) Lack of a single unifying problem; (iii) insufficient skill sets; (iv) traditional scientific approach; and (v) absence of investment. These challenges are natural yet concerning. It is believed that with time and much-advanced machine learning approaches can address these challenges [92].

Automation by AI is a burning issue as it will lead to unemployment in a large number of populations [96]. In exports view, the AI that is being used artificial narrow intelligence ANI is not really up to replace humans but will enhance humans and their laziness [35]. With AI, a person can save a lot of time and can use turn that saves time to creativity. Especially in the rational drug discovery process, the robots or machines can compile the data by understanding the subject matter. They can filter those compiled data and present them to the scientist, where the scientists or researchers will have to think on bigger

pictures of that study without worrying about the data and wasting time compiling them. The robots or machines will help the drug designer scientist design better, greener chemicals, and the synthesis process can be easier. At every point the machine can monitor and verify the process and data to minimize mistakes.

Declarations

Conflict of Interest: No potential conflict of interest is being reported by the authors.

References

[1] Renz P, Hochreiter S, Klambauer G (2019). Uncertainty estimation methods to support decision-making in early phases of drug discovery. In: *Workshop on Safety and Robustness in Decision-making* at 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[2] Berdigaliyev N, Aljofan M (2020). An overview of drug discovery and development. *Future Med Chem*; 12(10):939-947. [CrossRef]

[3] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, *et al.* (2017). Mastering the game of go without human knowledge. *Nature*; 550:354-359. [CrossRef]

[4] Hsu FH (2002). Behind deep blue: building the computer that defeated the world chess champion. Princeton University Press.

[5] Bojarski M, Testa DD, Dworakowski D, Firner B, Flepp B, *et al.* (2016). End to end learning for self-driving cars. *arXiv*:1604.07316

[6] Alam I, Khusro S, Khan M (2021). Personalized content recommendations on smart tv: challenges, opportunities, and future research directions. *Entertainment Comput*; 38:100418. [CrossRef]

[7] Colombo AW, Karnouskos S, Yu X, Kaynak O, Luo RC, *et al.* (2021). A 70-year industrial electronics society evolution through industrial revolutions: the rise and flourishing of information and communication technologies. *IEEE Industrial Electronics Magazine*; 15(1):115-126. [CrossRef]

[8] Wu H, Yin H, Chen H, Sun M, Liu X, Yu Y, Tang Y, *et al.* (2021). A deep learning–based smartphone platform for cutaneous lupus erythematosus classification assistance—simplifying the diagnosis of complicated diseases. *J Am Acad Dermatol*; S0190-9622(21):00402-03. [CrossRef] [PubMed]

[9] Undey C (2021). AI in process automation. *SLAS Technol*; 26(1):1-2. [CrossRef]

[10] Malandraki-Miller S, Riley PR (2021). Use of artificial intelligence to enhance phenotypic drug

discovery. *Drug Discov Today*; 26(4):887-901. [CrossRef]   [PubMed]

[11] Green CP, Engkvist O, Pairaudeau G (2018). The convergence of artificial intelligence and chemistry for improved drug discovery. *Future Med Chem*; 10(22):2573-2576. [CrossRef]

[12] Poduri R (2021). Historical Perspective of Drug Discovery and Development. In: Poduri R (eds) *Drug Discovery and Development.* Springer, Singapore: p. 1-10. [CrossRef]

[13] Jordan AM (2018). Artificial intelligence in drug design—the storm before the calm?. *ACS Med Chem Lett*; 9(12):1150-1152. [CrossRef]   [PubMed]

[14] Folmer RHA (2016). Integrating biophysics with HTS-driven drug discovery projects. *Drug Discov Today*; 21(3):491-498. [CrossRef]   [PubMed]

[15] Liu R, Li X, Lam KS (2017). Combinatorial chemistry in drug discovery. *Curr Opin Chem Biol*; 38:117-126. [CrossRef]   [PubMed]

[16] Aldewachi H, Al-Zidan RN, Conner MT, Salman MM (2021). High-throughput screening platforms in the discovery of novel drugs for neurodegenerative diseases. *Bioengineering (Basel)*; 8(2):30. [CrossRef]   [PubMed]

[17] Colombo M, Peretto I (2008). Chemistry strategies in early drug discovery: an overview of recent trends. *Drug Discov Today*; 13(15-16):677-684. [CrossRef]   [PubMed]

[18] Kraus JL (2021). From 'molecules of life' to new therapeutic approaches, an evolution marked by the advent of artificial intelligence: the cases of chronic pain and neuropathic disorders. *Drug Discov Today*; 26(4):1070-1075. [CrossRef]   [PubMed]

[19] Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, *et al.* (2021). A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *arXiv*:2102.10062.

[20] Kim SK, Huh JH (2020). Artificial neural network blockchain techniques for healthcare system: focusing on the personal health records. *Electronics*; 9(5):763. [CrossRef]

[21] Wang C, O'Neill Sm, Rothrock N, Gramling R, Sen A, *et al.* (2009). Comparison of risk perceptions and beliefs across common chronic diseases. *Prev Med*; 48(2):197-202. [CrossRef]   [PubMed]

[22] De Rycker M, Baragaña B, Duce SL, Gilbert IH (2018). Challenges and recent progress in drug discovery for tropical diseases. *Nature*; 559:498-506. [CrossRef]

[23] Bhutani P, Joshi G, Raja N, Bachhav N, Rajanna PK, *et al.* (2021). U.S. FDA approved drugs from 2015–June 2020: a perspective. *J Med Chem*; 64(5):2339-2381. [CrossRef]   [PubMed]

[24] Mak KK, Pichika MR (2018). Artificial intelligence in drug development: present status and future prospects. *Drug Discov Today*; 24(3):773-780. [CrossRef]

[25] Scannell JW, Blanckley A, Boldon H, Warrington B (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov*; 11:191-200. [CrossRef]

[26] Ma CKK, Danta M, Day R, Ma DDF (2018). Dealing with the spiralling price of medicines: issues and solutions. *Intern Med J*; 48(1):16-24. [CrossRef]   [PubMed]

[27] Zhavoronkov A, Vanhaelen Q, Oprea TI (2020). Will artificial intelligence for drug discovery impact clinical pharmacology?. *Clin Pharmacol Ther*; 107(4):780-785. [CrossRef] [PubMed]

[28] Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH (2017). Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic Reviews*; 6:245. [CrossRef]

[29] Hu J (2021). A new era for AI HPC and IC technologies in the transition to an intelligent digital world. In: *Metrology, Inspection, and Process Control for Semiconductor Manufacturing XXXV*; 11611(03) at SPIE Advanced Lithography. [CrossRef]

[30] Gawehn E, Hiss JA, Schneider G (2016). Deep learning in drug discovery. *Mol Inform*; 35(1):3-14. [CrossRef]   [PubMed]

[31] Lavecchia A (2019). Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discov Today*; 24(10):2017-2032. [CrossRef]   [PubMed]

[32] Fotis C, Antoranz A, Hatziavramidis D, Sakellaropoulos T, Alexopoulos LG (2018). Network-based technologies for early drug discovery. *Drug Discov Today*; 23(3):626-635. [CrossRef]   [PubMed]

[33] Lavecchia A (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*; 20(3):318-331. [CrossRef]

[34] Yang X, Wang Y, Byrne R, Schneider G, Yang S (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev*; 119(18):10520-10594. [CrossRef]   [PubMed]

[35] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK (2020). Artificial intelligence in drug discovery and development. *Drug Discov Today*; 26(1):80-93. [CrossRef]   [PubMed]

[36] Rai S, Raj U, Tichkule S, Kumar H, Mishra S, *et al.* (2016). Recent trends in in-silico drug discovery. *Int. J Comput Biol*; 5(1):54-76.

[37] Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, *et al.* (2020). Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent Sci*; 6(6):939-949. [CrossRef] [PubMed]

[38]   Hu SS, Chen P, Gu P, Wang B (2020). A deep learning-based chemical system for QSAR prediction. *IEEE J Biomed Health Inform*; 24(10):3020-3028. [CrossRef]   [PubMed]

[39]   Guedes IA, Pereira FSS, Dardenne LE (2018). Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front Pharmacol*; 9:1089. [CrossRef] [PubMed]

[40]   Stumpfe D, Bajorath J (2020). Current trends, overlooked issues, and unmet challenges in virtual screening. *J Chem Inf Model*; 60(9):4112-4115. [CrossRef]

[41]   Rifaioglu AS, Atas H, Martin MJ, *et al.* (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform*; 20(5):1878-1912. [CrossRef]   [PubMed]

[42]   Wang C, Xu P, Zhang L, *et al.* (2018). Current strategies and applications for precision drug design. *Front Pharmacol*; 9:787. [CrossRef] [PubMed]

[43]   Batool M, Ahmad B, Choi SJ (2019). A structure-based drug discovery paradigm. *Int J Mol Sci*; 20(11):2783. [CrossRef]   [PubMed]

[44]   Miyao T, Kaneko H, Funatsu K (2016). Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J Chem Inf Model*; 56(2):286-299. [CrossRef]   [PubMed]

[45]   Wang X, Song K, Li L, Chen L (2018). Structure-based drug design strategies and challenges. *Curr Top Med Chem*; 18(12):998-1006.   [CrossRef]   [PubMed]

[46]   Zhong F, Xing J, Li X, *et al.* (2018). Artificial intelligence in drug design. *Sci China Life Sci*; 61(10):1191-1204. [CrossRef]

[47]   Gupta A, Müller AT, Huiman BJH, *et al.* (2018). Generative recurrent networks for de novo drug design. *Mol Inform*; 37(1-2):1700111. [CrossRef] [PubMed]

[48]   Hessler G, Baringhaus KH (2018). Artificial intelligence in drug design. *Molecules*; 23(10):2520. [CrossRef]

[49]   Segler MHS, Kogej T, Tyrchan C, Waller MP (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci*; 4(1):120-131. [CrossRef]   [PubMed]

[50]   Popova M, Isayev O, Tropsha A (2018). Deep reinforcement learning for de novo drug design. *Sci Adv*; 4(7):eaap7885. [CrossRef]

[51]   Ertl P, Lewis R, Martin E, Polyakov V (2017). In silico generation of novel, drug-like chemical matter using the LSTM neural network. *arXiv*:1712.07449.

[52]   Yasonik J (2020). Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. *J Cheminform*; 12:14. [CrossRef]

[53]   Karim MR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S (2019). Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. *arXiv*: 1908.01288.

[54]   Kadurin A, Nikolonko S, Khrabrov K, Aliper A, Zhavoronkov A (2017). druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico.   *Mol Pharm*; 14(9):3098-3104.   [CrossRef]   [PubMed]

[55]   Brown N, Ertl P, Lewis R, Luksch T, *et al.* (2020). Artificial intelligence in chemistry and drug design. *J Comput Aided Mol Des*; 34:709-705. [CrossRef]

[56]   Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A (2017). Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *ChemRxiv Preprint*. [CrossRef]

[57]   Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, *et al.* (2018). Reinforced adversarial neural computer for de novo molecular design. *J Chem Inf Model*; 58(6):1194-1204. [CrossRef] [PubMed]

[58]   De Cao N, Kipf T (2018). MolGAN: an implicit generative model for small molecular graphs. *arXiv*:1805.11973

[59]   Maziarka Ł, Pocha A, Kaczmarczyk J, *et al.* (2020). Mol-CycleGAN: a generative model for molecular optimization. *J Cheminform*; 12:2. [CrossRef]

[60]   Khemchandani Y, O'Hagan S, Samanta S, *et al.* (2020). DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach. *J Cheminform*; 12:53. [CrossRef]

[61]   Druzhilovskiy DS, Stolbov L, Savosina P, *et al.* (2020). Computational approaches to identify a hidden pharmacological potential in large chemical libraries. *J Supercomput Front Innov*; 7. [CrossRef]

[62]   Born J, Manica M, Cadow J, *et al.* (2021). Data-driven molecular design for discovery and synthesis of novel ligands-a case study on SARS-CoV-2. *Mach Learn Sci Technol*; 2(2):025024. [CrossRef]

[63]   Mughal H, Wang H, Zimmerman M, *et al.* (2021). Random forest model prediction of compound oral exposure in the mouse. *ACS Pharmacol Transl Sci*; 4(1):338-343. [CrossRef]

[64]   Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF (2017). Convolutional embedding of attributed molecular graphs for physical property prediction. *J Chem Inf Model*; 57(8):1757-1772. [CrossRef]   [PubMed]

[65] Conan M, Théret N, Langouet S, Siegel A (2021). Constructing xenobiotic maps of metabolism to predict the role of enzymes in DNA adduct formation. *Research Square*. [CrossRef]

[66] Coley CW, Barzilay R, Jaaakkola TS, Green WH, Jensen KF (2017). Prediction of organic reaction outcomes using machine learning. *ACS Cent Sci*; 3(5):434-443. [CrossRef]

[67] Chen X, Li H, Tian L, Li Q, Luo J, Zhang Y (2020). Analysis of the physicochemical properties of acaricides based on Lipinski's rule of five. *J Comput Biol*; 27(9):1397-1406. [CrossRef]   [PubMed]

[68] Shultz MD (2019). Two decades under the influence of the rule of five and the changing properties of approved oral drugs. *J Med Chem*; 62(4):1701-1714. [CrossRef]

[69] Ley SV, Fitzpatrick DE, Ingham RJ, Myers RM (2015). Organic synthesis: march of the machines. *Angew Chem Int Ed*; 54(11):3449-3464. [CrossRef]

[70] Zhang W, Cue BW (2018). Green techniques for organic synthesis and medicinal chemistry. Wiley Online Library. [CrossRef]

[71] Davies IW (2019). The digitization of organic synthesis. *Nature*; 570(7760):175-181. [CrossRef]

[72] Gromski PS, Granda JM, Cronin L (2019). Universal chemical synthesis and discovery with 'the chemputer'. *Trends Chem*; 2(1):4-12. [CrossRef]

[73] Johansson S, Thakkar A, Kogej T, *et al.* (2019). AI-assisted synthesis prediction. *Drug Discov Today Technol*; 32-33:65-72. [CrossRef]   [PubMed]

[74] Steiner S, Wold J, Glatzel S, *et al.* (2019). Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*; 363(6423):eaav2211. [CrossRef]

[75] Wang Z, Zhao W, Hao G, Song B (2021). Mapping the resources and approaches facilitating computer-aided synthesis planning. *Org Chem Front*; 8(4):812-824. [CrossRef]

[76] Orloff J, Douglas F, Pinheiro J, Levinson S, *et al.* (2009). The future of drug development: advancing clinical trial design. *Nat Rev Drug Discov*; 8(12):949-957. [CrossRef] [PubMed]

[77] Feijoo F, Palopoli M, Bernstein J, *et al.* (2020). Key indicators of phase transition for clinical trials through machine learning. *Drug Discov Today*; 25(2):414-421. [CrossRef]

[78] Ménard T, Koneswarakantha B, Rolo D, Barmaz Y, Popko L, Bowling R (2020). Follow-up on the use of machine learning in clinical quality assurance: can we detect adverse event under-reporting in oncology trials?. *Drug Saf*; 43:295-296. [CrossRef]

[79] Zame WR, Bica I, Shen C, Curth A, *et al.* (2020). Machine learning for clinical trials in the era of COVID-19. *Statis Biopharma Res*; 12(4):506-517. [CrossRef]

[80] Levin JM, Oprea TI, Davidovich S, Clozel T, Overington JP, *et al.* (2020). Artificial intelligence, drug repurposing and peer review. *Nat Biotechnol*; 38(10):1127-1131. [CrossRef]   [PubMed]

[81] Tanoli Z, Vähä-Koskela M, Aittokallio T (2021). Artificial intelligence, machine learning and drug repurposing in cancer. *Expert Opin Drug Discov*; 1-13.   [CrossRef]   [PubMed]

[82] Luo H, Li M, Yang M, Wu FX, Li Y, Wang J (2021). Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief Bioinform*; 22(2):1604-1619. [CrossRef] [PubMed]

[83] Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, *et al.* (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun*; 8(1):573. [CrossRef]   [PubMed]

[84] Dissez G, Ceddia G, Pinoli P, Ceri S, Masserolli M (2019). Drug repositioning predictions by non-negative matrix tri-factorization of integrated association data. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*; 2019:25-33. [CrossRef]

[85] Zheng S, Ma H, Wang J, Li J (2019). A computational bipartite graph-based drug repurposing method. In: Vanhaelen Q (Ed). *Methods in Molecular Biology; 1903: Computational Methods for Drug Repurposing*. Humana Press, New York: pp 115-127. [CrossRef]

[86] Shahreza ML, Ghadiri N, Mousavi SR, Varshosaz J, Green JR (2018). A review of network-based approaches to drug repositioning. *Brief Bioinform*; 19(5):878-892. [CrossRef]   [PubMed]

[87] Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*; 35(24):5191-5198.   [CrossRef] [PubMed]

[88] Smith JS, Roitberg AE, Isayev O (2018). Transforming computational drug discovery with machine learning and AI. *ACS Med Chem Lett*; 9(11):1065-1069. [CrossRef]   [PubMed]

[89] Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S (2019). Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci*; 40(8):592-604. [CrossRef]   [PubMed]

[90] Jiménez-Luna J, Grisoni F, Schneider G (2020). Drug discovery with explainable artificial intelligence. *Nat Mach Intell*; 2:573-584. [CrossRef]

[91] Benhenda M (2017). ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? *arXiv*:1708.08227

[92]   Henstock PV (2019). Artificial intelligence for pharma: time for internal investment. *Trends Pharmacol Sci*; 40(8):543-546. [CrossRef]

[93]   Iolanda Bulgaru (2021). Pharma industry in the age of artificial intelligence: the future is bright. *Healtcare Weekly*. https://healthcareweekly.com/artificial-intelligence-in-pharmacology/ (accessed 10 June 2021).

[94]   IBM News Room (2016). IBM and Pfizer to accelerate immuno-oncology research with Watson for drug discovery. https://newsroom.ibm.com/2016-12-01-IBM-and-Pfizer-to-Accelerate-Immuno-oncology-Research-with-Watson-for-Drug-Discovery (accessed 10 June 2021).

[95]   XtalPi Inc. (2018). Announces strategic research collaboration with Pfizer Inc. to develop artificial intelligence-powered molecular modeling technology for drug discovery. In: Cision PR Newswire. https://www.prnewswire.com/news-releases/xtalpi-inc-announces-strategic-research-collaboration-with-pfizer-inc-to-develop-artificial-intelligence-powered-molecular-modeling-technology-for-drug-discovery-300644351.html (accessed 10 June 2021).

[96]   Wolters L (2020). Robots, automation, and employment: where we are. *MIT Work of the Future: Working Paper Series.* https://workofthefuture.mit.edu/research-post/robots-automation-and-employment-where-we-are/ (accessed 10 June 2021).